Technical white paper

HP ProLiant DL980 G7を用いた大規模創薬モデリング　報告書

生命における共通の最小単位である細胞は、外部刺激によって大きく影響を受ける。その刺激を受ける蛋白質の一つは、G protein-coupled receptor（GPCR）である。GPCRは細胞膜を貫通して細胞内外に局在し、細胞外側で刺激を受けると細胞内側へ信号を流す。
その結果、細胞は正常状態と異常状態の間を遷移する。異常状態の細胞のGPCRに、化合物と相互作用をさせて正常状態に戻す可能性があることが分かってきた現在、効率的に薬効を与える化合物を見つけることが大きな課題になっている。限られている資源と労力の中、計算機の力で活性の高い創薬リード化合物を発見することに注目が集まっている。計算機による化合物スクリーニング法として、化合物間の類似度行列と蛋白質間の類似度行列から、相互作用の可能性が高い組み合わせを推定する高精度のアルゴリズムが存在するが、アルゴリズムの中に構築されるモデルの計算量が増大するため、通常のワークステーションや中型サーバーでは使用できる相互作用データ数に限界がある。
また、シングルCPUで大量の化合物データの類似度を計算すると、計算時間が膨大になってしまう。Python言語のmultiprocessingライブラリで化合物類似度計算の並列化と、C++言語のOpenMP拡張をSupport Vector Machine学習アルゴリズムに適用することによって、インテル® Xeon® プロセッサーE7 ファミリー8基（80コア）を搭載した、HP ProLiant DL980 G7のTB（テラバイト）単位メモリ構成で大型創薬相互作用モデルの構築を狙う研究が立ち上がりつつある。
このホワイトペーパーでは、研究を担当する京都大学大学院薬学研究科システム創薬科学講座博士研究員J.B.Brown 氏が行ったベンチマーク結果をご紹介する。

1. 背景

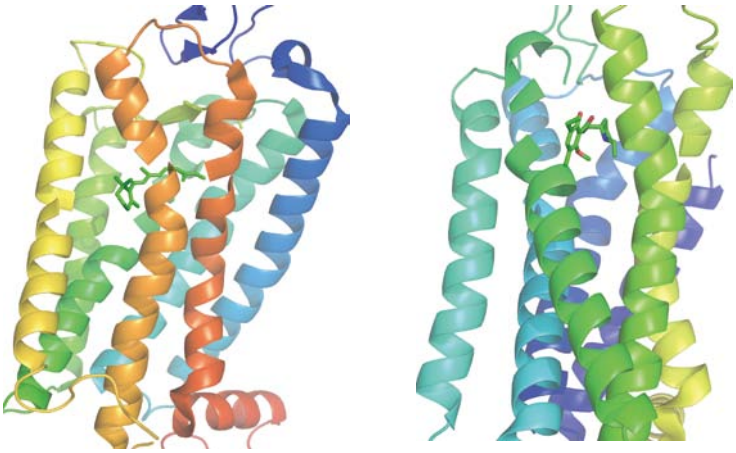
1.1 GPCRの構造と重要性

GPCRは細胞膜を貫通する7つのヘリックスと、ヘリックスを繋げるループ領域から構成され、小さな光の粒子から大きな化合物まで幅広い分子と相互作用を行います。図1は、2000年に構造が解明されたロドプシン（rhodopsin）というGPCR（*Science* 289:739）と、2010年に解明されたドーパミン（dopamine）というGPCR（*Science* 330:1091）のそれぞれの構造を示しており、貫通領域においてヘリックスがGPCR構造を形成する特徴が明らかです。ロドプシンは目に存在し、貫通領域に共存する補因子の立体構造変化に伴って光の強さを伝えます。ドーパミンは神経細胞系で、情報を伝達する役割を果たします。この二つを含めて約1000種類のGPCRがあります。

intel inside
Xeon
インテル® Xeon® プロセッサー
E7 ファミリー

基本的に各GPCRに結合する補因子は存在しますが、補因子の代わりにGPCRに別の化合物が結合するとGPCRの立体構造と機能に変化が与えられます。その変化によって、下流の信号伝達が変わって細胞に影響を与えます。相互作用から細胞への影響のプロセスが広く知られているため、異常状態の細胞の将来的な治療方法として、GPCRに結合して薬効を与える新薬が期待されています。そのため、創薬研究の発展がますます重要になると考えられます。

図1. （左）2000年にGPCRとして世界初めて結晶構造が解明されたrhodopsin（PDB: 1F88）。（右）2010年に解明されたdopamine D_3（PDB: 3PBL）。細胞膜を貫通する7つのヘリックスが特徴である。上：細胞外、下：細胞内



1.2 計算機による化合物スクリーニング

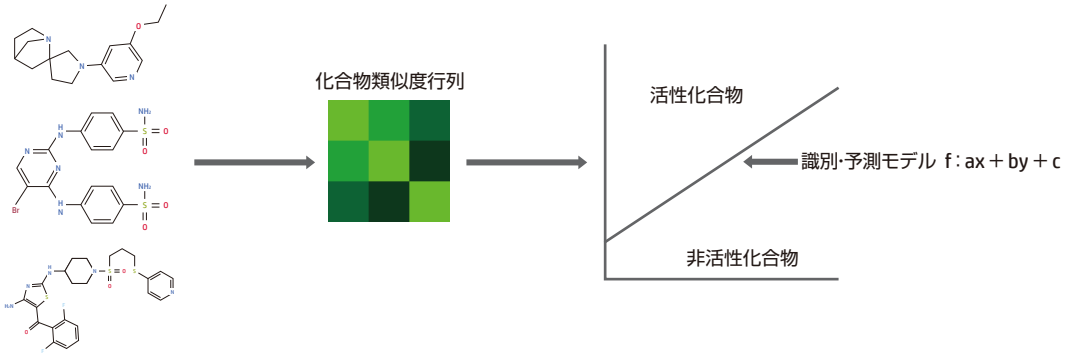
近年、創薬の研究は、計算科学の最新技術を取り入れる形で、大規模分子シミュレーションと化合物カタログのインシリコ·スクリーニングに至っています。10年前にクラスター計算機を要していた計算は今、安価な個人用ワークステーションでもできるようになりました。しかしながら、図1のような結晶構造が解明されているGPCRは、現時点では極めて少数です。また、創薬において大きな役割を果たす有機化合物の数は10^{60}程度存在していると言われており、その化合物の中からごく一部を選択して高精度のシミュレーションを行うことは、未だ困難です。従って結合シミュレーションができない場合の対応が必要であり、次の二つの手法が提案されています。

2. 計算による次世代創薬アプローチ

2.1 標的が一つの場合

標的タンパク質が一つに限定される場合は、定量的構造活性相関解析（QSAR）がよく使われます。QSARは化合物の定量的な特徴から、実験的な活性を予測するモデル関数を算出する手法です。QSARのモデル構築方法の一つとして、サポートベクターマシン（SVM, Cortes and Vapnik, *Machine Learning* 20:273）という高性能機械学習アルゴリズムがあります。SVMを用いるQSARは化合物同士の類似度行列を入力とし、例えば活性値が10μM以下である化合物とそうでない化合物を区別する最適モデル関数を計算します。そして活性が未知な化合物とモデル算出に用いた化合物との類似度を計算してSVMモデルに入力すると、活性有無の予測判定ができます。このプロセスは図2で示されています。

図2. 学習用の化合物の特徴情報から類似度行列を計算して、活性情報と共に機械学習アルゴリズムに入力します。活性を予測するモデルが算出され、活性が未知な化合物に適用して活性有無を予測できます。



2.2 複数標的の場合

図2のQSAR方法は、既知の活性データに基づいてモデル算出を行いますが、結合データの少ない、又は結合データの無いGPCRが少なくありません。もう一方で、化合物が各蛋白質によって異なる活性を示すため、特定の標的蛋白質活性情報を別の蛋白質QSAR解析に適用することができません。従って、既存の薬が別の病気に対しても薬効があるかどうかをQSARだけで調べることは不可能です。

このような背景から、近年、数千程度の大量の化合物と蛋白質データを用いて、化合物と蛋白質の相互作用を一括にして解釈する挑戦的な研究が始まっています。この手法は図2の化合物類似度行列と同様な蛋白質の類似度行列を合わせて、相互作用のモデルを構築します。算出されるモデルを適用すると、特定の標的蛋白質に対して新しい化合物の相互作用を予測することができたり、創薬シーズ化合物を早期に特定できたりします。図3にこの創薬のアプローチ（chemical genomics-based virtual screening、略CGBVS）が示されています。

図3. 大量のケミカルゲノミックスのデータを活かすCGBVS法

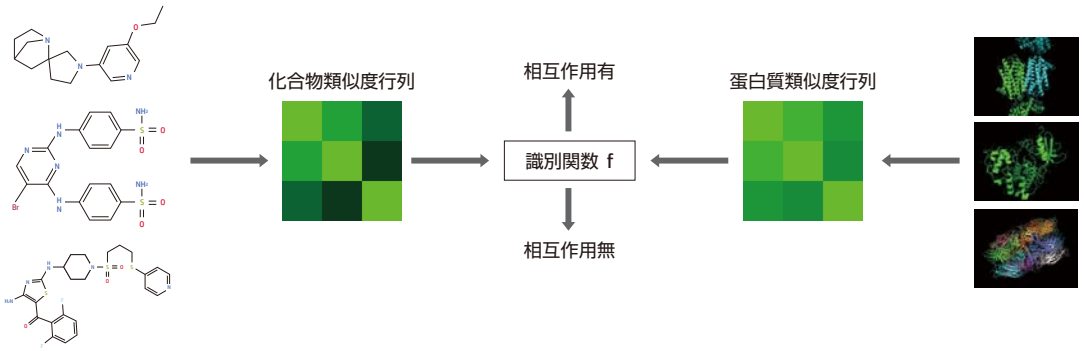


図3のCGBVS方の一つの利点は、自然リガンドや補因子の知られていない蛋白質（orphan）に対しても、特徴が似ている蛋白質とその蛋白質の相互作用化合物データから、結合する可能性の高い化合物を予測することができることです。そしてorphanに結合する化合物を発見できれば、そのGPCRが細胞においてどのような役割を果たしているか手がかりを得ることができます。結果として新薬の候補を得られるのみならず、細胞組織と機能の理解を更に深めることができます。

3. 従来の結果

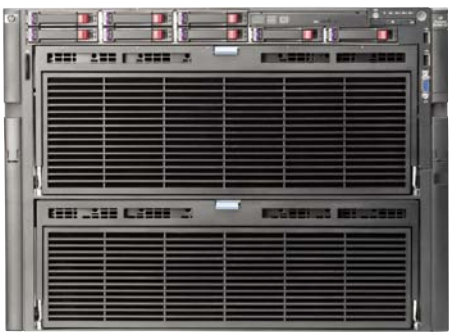
有機化合物空間の大きさは10^{60}であるため、QSARとCGBVSの学習データとしては一部の化合物しか使用できません。検証実験が行われたYabuuchi et al., *Mol Syst Biol* 7:472（2011）の論文では、学習例数が5200でした。また、Brown et al., *BIBM PIDD* 136（2012）の論文では、学習例数を約2倍の10000まで増やして相互作用モデルを計算しました。

4. HP ProLiant DL980 G7による大規模計算

4.1 高速化する方法

HP ProLiant DL980 G7を用いる実験を2種類行いました。まず、QSARの化合物類似度行列の計算を行い、通常のワークステーションの計算速度と比較しました。そして相互作用の予測モデル構築に用いる学習例数を大幅に増加し、SVM学習アルゴリズムが計算を終了して最適なモデルを算出できるか実験を行いました。類似度行列計算の並列化により計算時間を減らせることが考えられます。化合物類似度計算を行うために、Python言語のmultiprocessingモジュールを利用し、以下のように計算を並列化しました：

```
import multiprocessing

def computeWorker( data, worker_id, totalWorkers,
                   outputFile, file_lock,
                   **kwargs ):

    def writeBuffer( filename, fileLock, bufferData ):
        fileLock.acquire()
        ... # bufferData をディスクに書き込む作業
        fileLock.release()
    ## データを保存する関数終了

    ... # 初期化

    try:
        ... # ループカウンターを更新
        if counter % totalWorkers == worker_id: # このプロセスで計算する
            ... # 実際の計算と一時保存
            if len( buffer ) == bufferLimit:
                writeBuffer( outputFile, file_lock, buffer )
        ... # その他処理
    except Exception, msg:
        logger.critical( "Exception がありました―%s:%s" % ( workerName, msg ) )
## 1 プロセスで動く関数終了


def masterProcess( maxWorkers = None, ... ): # 引数省略

    ... # 事前処理

    # CPU 数をもとに起動するプロセスを設定する
    numWorkers = multiprocessing.cpu_count() if maxWorkers is None else int(maxWorkers)
    # 出力ファイルのロックを作る
    useFileLock = multiprocessing.Lock()

    # 並列プロセスを作る
    processPool = list()
    for worker_counter in xrange( numWorkers ):
        newProcess = multiprocessing.Process(
                     target = computeWorker,
                     name = 'computeWorker-%i' % ( worker_counter + 1 ),
                     args = ( data, worker_counter, numWorkers, dbFilename, useFileLock ),
                     kwargs = kwargs )
        processPool.append( newProcess )

    # 並列プロセスを開始して終了を待つ
    for p in processPool: p.start()
    for p in processPool: p.join()

    ... # 計算後処理

## 並列プロセスを生成する関数終了
```



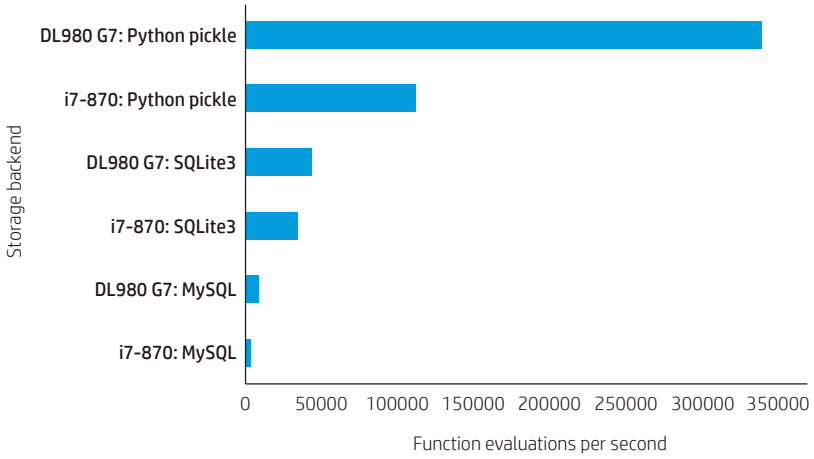
ハードウェア：**HP ProLiant DL980 G7**
搭載プロセッサー：インテル® Xeon® プロセッサー E7-4870 2.4GHz 8基／80コア
搭載メモリー：2TB
搭載記憶領域：900GB SAS HDD×8台　合計7.2TB
OS：Red Hat Enterprise Linux 6.2

インテル® Xeon® プロセッサー
E7 ファミリー

4.2 並列計算による加速化

1個のCPUで類似度行列計算を行う場合は、同時に計算と出力をすることが可能ですが、並列計算の場合は特定の計算順番がないため、同時出力が不可能で、計算結果を保存する措置が必要です。今回は、類似度関数の結果を保存する3通りのバックエンドを用いて、Intel i7-870のワークステーションとHP ProLiant DL980 G7の並列処理速度を計りました。実験結果は図4に示されています。

図4. Intel i7-870ワークステーションとDL980 G7の並列化による化合物類似度計算速度。単位：類似度関数/秒



4.3 大量RAMを用いた大規模相互作用モデルの構築

QSARに基づく処理の計算時間比較で、HP ProLiant DL980 G7の高性能が確認できたため、最後に相互作用予測モデルの構築においてHP ProLiant DL980 G7がどの程度の学習例数に対応できるか実験を行いました。最初に行った実験では100,000例数を入力して学習させました。この学習問題はBrown et al. *BIBM PIDD* 136（2012）の相互作用データの10倍に相当します。C++のOpenMP拡張をLibSVM（Chang and Lin, *ACM Trans Intell Sys Tech* 2:27）に応用し、HP ProLiant DL980 G7の全80コアを学習に使った結果、一日以内にモデルの構築ができました。

10倍の学習データを用いた実験に成功したことから、最後にさらに10倍の学習例数（1,000,000）の実験に挑戦しました。学習例数が増加されれば増加されるほど計算量が二乗的に増えるのを考慮すると、100万の学習例数実験の計算が短期間に終了するとは予期できませんでした。しかし、HP ProLiant DL980 G7の並列計算に適している処理性能によって、SVMアルゴリズムが、無事に実験期間内で100万の学習例に基づく相互作用モデルの算出に成功しました（要計算期間：14 日）。通常のワークステーションや中型サーバーでは考えられない出来事でした。

5. 今後の活用

最初に述べたようにGPCRは1000種類があり、全GPCRと10^{60}化合物の全組み合わせを学習することが不可能ですが、HP ProLiant DL980 G7の利用によって、100万程度の相互作用サンプルからモデルを構築することが可能であることが分かりました。今後、100万程度のモデルを如何に組み合わせるかを検討の上、GPCRにおける相互作用を、より包括的に解釈するモデルを目指したいと考えています。また、GPCR以外の創薬標的も存在するため、その標的についても同様な計算並列化による解析を行うことが重要だと考えられます。

安全に関するご注意　ご使用の際は、商品に添付の取扱説明書をよくお読みの上、正しくお使いください。水、湿気、油煙等の多い場所に設置しないでください。火災、故障、感電などの原因となることがあります。

お問い合わせはカスタマー・インフォメーションセンターへ
03-5749-8328 月～金 9:00～19:00　土 10:00～17:00（日、祝祭日、年末年始および5/1を除く）
機器のお見積もりについては、代理店、または弊社営業にご相談ください。
HP ProLiantに関する情報は http://www.hp.com/jp/proliant

Intel、インテル、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Xeon、Xeon Insideは、アメリカ合衆国および / またはその他の国におけるIntel Corporationの商標です。
記載されている会社名および商品名は、各社の商標または登録商標です。
記載事項は2013年3月現在のものです。
本カタログに記載されている情報は取材時におけるものであり、閲覧される時点で変更されている可能性があります。あらかじめご了承ください。
© Copyright 2013 Hewlett-Packard Development Company,L.P.

本カタログは、環境に配慮した用紙と植物性大豆油インキを使用しています。 PRINTED WITH SOY INK

日本ヒューレット・パッカード株式会社
〒136-8711 東京都江東区大島2-2-1